732A38 Computational statistics
Division of Statistics and Machine Learning
Department of Computer and Information Science

Computer lab 5

Instructions

- Create a report to the lab solutions in PDF.
- Be concise and do not include unnecessary printouts and figures produced by the software and not required in the assignments.
- **Include all your codes as an appendix into your report.**
- A typical lab report should 2-4 pages of text plus some amount of figures plus appendix with codes.
- The lab report should be submitted via LISAM before the deadline.

Assignment 1: Hypothesis testing

In 1970, the US Congress instituted a random selection process for the military draft. All 366 possible birth dates were placed in plastic capsules in a rotating drum and were selected one by one. The first date drawn from the drum received draft number one, the second date drawn received draft number two, etc. Then, eligible men were drafted in the order given by the draft number of their birth date. In a truly random lottery there should be no relationship between the date and the draft number. Your task is to investigate whether or not the draft numbers were randomly selected. The draft numbers (Y=Draft_No) sorted by day of year (X=*Day_of_year*) are given in the file **lottery.xls**

1. Make a scatterplot of Y versus X and conclude whether the lottery looks random.
2. Compute an estimate $\tilde{Y}$ of the expected response as a function of X by using a loess smoother (use loess()), put the curve $\tilde{Y}$ versus X in the previous graph and state again whether the lottery looks random.
3. To check whether the lottery is random, it is reasonable to use test statistics
$$T = \frac{\tilde{Y}\{(X_b) - \tilde{Y}(X_a)}{X_b - X_a}, where\ X_b = \arg\max_X \tilde{Y}, X_a = \arg\min_X \tilde{Y}.$$
If this value is significantly greater than zero, then there should be a trend in the data and the lottery is not random. Estimate the distribution of T by using a non-parametric bootstrap with $B = 2000$ and comment whether the lottery is random or not. What is the P-value of the test?
4. Implement a function depending on *data* and *B* and that tests the hypothesis $H_o: Lottery\ is\ random$ vs $H_a: Lottery\ is\ not\ random$ by using a permutation test with statistics T and returns the p-value of this test. Test this function on our data with *B=2000*.
5. Make a crude estimate of the power of the test constructed in step 4:

a. Generate (an obviously non-random) dataset with $n = 366$ observations by using same X as in the original data set and $Y(x) = \max(0, \min(\alpha x + \beta, 366))$ where $\alpha = 0.1$ and $\beta \sim N(183, sd = 10)$
b. Plug these data into the permutation test with $B = 200$ and note whether it was rejected
c. Repeat steps a)-b) for $\alpha = 0.2, 0.3, \ldots, 10$

What can you say about the quality of your test statistics considering the value of the power?

Assignment 2: Bootstrap, jackknife and confidence intervals

The data you are going to continue analyzing is the database of home prices in Albuquerque, 1993. The variables present are Price; SqFt – the area of a house; FEATS – number of features such as dishwasher, refrigerator and so on; Taxes – annual taxes paid for the house. Explore the file ***prices1.xls***

1. Plot the histogram of Price. Does it remind any conventional distribution? Compute the mean price.
2. Estimate the distribution of the mean price of the house using bootstrap. Determine the bootstrap bias-correction and the variance of the mean price. Compute the 95% confidence interval for the mean price using bootstrap percentile, bootstrap BCa, and first-order normal approximation (**Hint**: use *boot(),boot.ci(),plot.boot(),print.bootci()*)
3. Estimate the variance of the mean price using the jackknife and compare it with the bootstrap estimate
4. Compare the confidence intervals obtained with respect to their length and the location of the estimated mean in these intervals.

Submission procedure

Assume that X is the current lab number.

If you are neither speaker nor opponent for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members submits the group report using *Lab X group report* in the *Submissions* folder before the deadline

If you are a speaker for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members does the following before the deadline:
    - submits the group report using *Lab X group report* in the *Submissions* folder before the deadline
    - Goes to Study room *Speakers X→Documents* and opens file *Password X.txt*. Then the student should put your group report into ZIP file *Lab X.zip* and protect it with a password you found in Password X.txt
    - Uploads the file to *Collaborative workspace* folder

If you are opponent for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members submits the group report using *Lab X group report* in the *Submissions* folder before the deadline
- After the deadline for the lab has passed, go to Collaborative workspace folder and download *Lab X.zip*. Open the PDF in this ZIP file by using the password available in *Course Documents→Password X.txt,* read it carefully and **prepare at least two questions/comments/improvement suggestions** in order to put them at the seminar (i.e. at least two questions per opponent)